Do we measure overconfidence? A closer look at the interval production task

Langnickel, Ferdinand ; Zeisberger, Stefan

Abstract: The most common test for overconfidence in the form of miscalibration—the Interval Production task (IP)—is based on the assumption that people internalize requested confidence levels. We demonstrate experimentally that decision makers' perceived confidence is, however, unaffected by variations in the requested confidence level. In addition, we find large heterogeneity in perceived confidence that the traditional IP measure fails to account for. We show that the alternative measure based on decision makers' perceived confidence by contrast yields coherent, moderate overconfidence levels. Our evidence suggests that the consistency of the two measures is limited and that they are related to different individual characteristics.

DOI: https://doi.org/10.1016/j.jebo.2016.04.019

Title: Do We Measure Overconfidence? A Closer Look at the Interval Production Task

Author: Ferdinand Langnickel Stefan Zeisberger

PII: S0167-2681(16)30076-2
DOI: http://dx.doi.org/doi:10.1016/j.jebo.2016.04.019
Reference: JEBO 3794

To appear in: *Journal of Economic Behavior & Organization*

Received date: 7-10-2015
Revised date: 23-4-2016
Accepted date: 28-4-2016

Please cite this article as: Ferdinand Langnickel, Stefan Zeisberger, Do We Measure Overconfidence? A Closer Look at the Interval Production Task, *<![CDATA[Journal of Economic Behavior and Organization]]>* (2016), http://dx.doi.org/10.1016/j.jebo.2016.04.019

Do We Measure Overconfidence? A Closer Look at the Interval Production Task*

Ferdinand Langnickel[a] and Stefan Zeisberger[†b]

[a]Department of Banking and Finance, University of Zurich, Plattenstrasse 32, 8032 Zurich, Switzerland

[b]College of Business, Stony Brook University, Stony Brook, NY 11794

April 22, 2016

Abstract

The most common test for overconfidence in the form of miscalibration—the Interval Production task (IP)—is based on the assumption that people internalize requested confidence levels. We demonstrate experimentally that decision makers' perceived confidence is, however, unaffected by variations in the requested confidence level. In addition, we find large heterogeneity in perceived confidence that the traditional IP measure fails to account for. We show that the alternative measure based on decision makers' perceived confidence by contrast yields coherent, moderate overconfidence levels. Our evidence suggests that the consistency of the two measures is limited and that they are related to different individual characteristics.

Keywords: overconfidence, miscalibration, methodology, experimental economics, experimental finance

JEL classification: D83, G02, G12

*We thank David Budescu, Thorsten Hens, Thomas Langer, Alexandra Niessen-Rünzi, Ulrich Schmidt, Karl Halvor Teigen, Leonard Walk, Martin Weber, and participants of the following seminars and conferences: Experimental Finance 2015 in Nijmegen, The Netherlands, SPUDM 2015 in Budapest, Hungary, Financial Economics Seminar at University of Zurich, Switzerland. This research was gratefully supported by the Swiss National Science Foundation (100018-149934).

†Corresponding author; email: stefan.zeisberger@stonybrook.edu

1 Introduction

Numerous studies have demonstrated that overconfidence can lead to suboptimal decisions in a variety of domains (Camerer and Lovallo, 1999; Johnson and Tierney, 2011; Malmendier and Tate, 2005; Odean, 1998). The most robust documentation of overconfidence is that people overestimate the precision of their own judgment (for a review, see Lichtenstein et al., 1982; Moore et al., 2014). This form of overconfidence, called miscalibration, overprecision, or judgmental overconfidence, induces people to rely too much on their own (biased) judgment and thus is an integral part of biased decision making in general (Bazerman and Moore, 2013).[1] The mainly used test for it is the Interval Production task (IP).[2] In the IP, decision makers are asked to provide lower and upper bound estimates (intervals) for a set of questions like "How long is the Nile river?". Subjects are instructed to state intervals such that their own confidence, that the true, unknown value falls between these stated bounds, equals a confidence level that is requested by the experimenter, for example 90%. On average, the ratio of true values that fall into decision makers' interval estimates, the "hit rate", should correspond to the requested confidence level (in this case 90%). However, commonly people are found to have much lower hit rates, so that they are classified as overconfident (Alpert and Raiffa, 1982; Russo and Schoemaker, 1992).

This paper aims to critically assess the suitability of the IP to measure overconfidence. Recent studies have already suggested that the IP measure may not function as presumed. In particular, it has been shown that groups with different requested confidence levels achieve the same average hit rate because they do not adjust the width of their interval estimates (Teigen and Jørgensen, 2005).[3] In addition, when estimating their own number of hits ("frequency judgments") people tend to be only conservatively confident in their own intervals compared to the typically high level of requested confidence (Cesarini et al., 2006; Winman et al., 2004). What remains unsolved, however, is the actual effect of the requested confidence level on the *perceived* level of confidence. Will people, for example, still provide conservative frequency judgments if the requested confidence level is low?

An experimental economist who studies the relation of overconfidence and economic behavior nonetheless can be agnostic about these points of critique as they address the

[1] In the following, we will use only the term overconfidence to refer to judgmental miscalibration, unless stated otherwise.

[2] See, e.g. Alpert and Raiffa (1982); Ben-David et al. (2013); Biais et al. (2005); Russo and Schoemaker (1992).

[3] These findings are based on between-subject variation of the requested confidence level. In response to within-subject variation of the requested confidence level decision makers have been shown to adjust their interval estimates (see, e.g. Alpert and Raiffa, 1982; Budescu and Du, 2007).

aggregate, not the individual level. Therefore, it is not surprising that the IP has been used in numerous experimental studies to elicit overconfidence at the individual level and link it to economic behavior (for most recent studies, see Ackert et al., 2015; Ben-David et al., 2013; Fellner-Röhling and Krügel, 2014; Herz et al., 2014), but the results are still surprisingly inconclusive overall.

Against this background, we address three open issues to help understanding the measurement of aggregate and individual overconfidence in the IP paradigm. First, one might wonder why decision makers are unable to adapt their intervals widths in the IP. To provide an answer we take a step back and test whether the requested confidence level has an effect on people's *perceived* level of confidence in the first place. Second, given the often unsuccessful experimental attempts to link overconfidence and economic behavior at the individual level, another important open question is whether the use of frequency judgments changes the relative ranking in overconfidence compared to the classic IP measure. Lastly, the two overconfidence measures may be explained by individual "background" characteristics such as cognitive abilities or the aversion of wide intervals which possibly explains some experimental results.

The IP overconfidence measure rests on the assumption that people adopt the requested confidence level and adjust their interval estimates accordingly. Using the classical IP paradigm we run an experiment in which we elicit *perceived confidence levels* in the form of frequency judgments, i.e. ex-post estimates of the number of hits in the IP, and we vary the degree of requested confidence. In two independent surveys in Switzerland and the U.S., we employ a between-subject design in which participants are randomly assigned to one of three treatments with requested confidence levels of $30\%, 60\%$ or 90%. We confirm weaknesses of the IP measure presented in Teigen and Jørgensen (2005) and show that decision makers not even adjust their frequency judgments to different levels of requested confidence. Using decision makers' frequency judgments, we find evidence that people respond to an individual confidence level that is unaffected by the requested confidence level. In all treatments, we observe large variations in individual confidence that the IP measure does not take into account. As a consequence, the consistency of the two overconfidence measures is limited. People might appear very overconfident in the IP simply because their true confidence level is overestimated by the requested level and vice versa. We conclude that the IP as it is currently used, i.e. comparing hit rates with requested confidence levels, has major shortcomings for measuring overconfidence at both the aggregate and individual level. As an alternative, we propose to use people's frequency

judgments to measure overconfidence.

We would like to mention that we are not the first to question the reliability of measuring overconfidence with the IP on the aggregate level. Next to the aforementioned studies (Teigen and Jørgensen, 2005; Cesarini et al., 2006; Winman et al., 2004), many previous studies have raised specific methodological concerns with consequences for the aggregate level of overconfidence. It has been shown, for example, that the alternative two-choice question format (e.g., Koriat et al., 1980) yields lower levels of overconfidence than the IP (Klayman et al., 1999). Similarly, exclusion instructions that induce people to think about the values that lie outside of their provided intervals in the IP yield smaller levels of overconfidence than inclusion questions (Soll and Klayman, 2004; Teigen and Jørgensen, 2005; Yaniv and Schul, 1997). Interestingly, other forms of overconfidence such as better-than-average beliefs (Svenson, 1981) or the illusion of control (Langer, 1975) were found to be inconsistent with the IP measure (Deaves et al., 2008; Glaser and Weber, 2007; Hilton et al., 2011; Menkhoff et al., 2006; Moore and Healy, 2008). Moreover, the level of overconfidence has been found to vary with the difficulty (Lichtenstein et al., 1982) and domain (Klayman et al., 1999) of the question set. Other studies attribute overconfidence to measurement errors and sampling effects of the question set that impair the validity of otherwise well-working cues (Gigerenzer et al., 1991; Juslin, 1993, 1994; Soll, 1996). Our work contributes to the understanding of how appropriate the IP measure is on the aggregate level.

Additionally, our findings can help to explain previous puzzling results on the individual level. While the IP has been found to predict some behavior like innovative activity, risk-taking and ordering decisions of managers (Ben-David et al., 2013; Herz et al., 2014; Ren and Croson, 2013), a wide range of studies is not able to explain economic behavior with the IP overconfidence measure. In a recent study, for example, Ackert et al. (2015) find that investors who diversify more in an experimental asset market appear more overconfident in the IP which is at odds with previous findings (e.g. Goetzmann and Kumar, 2008). Similarly, despite showing a negative relation of overconfidence and trading performance (Biais et al., 2005; Deaves et al., 2008; Kirchler, 2002) experimental studies mostly fail to confirm the empirically hypothesized link between overconfidence and trading volume (Biais et al., 2005; Fellner-Röhling and Krügel, 2014; Glaser and Weber, 2007; Kirchler, 2002); for empirical studies see (Barber and Odean, 2000, 2001; Statman et al., 2006).[4] These

[4]Deaves et al. (2008) do find a positive correlation between overconfidence and trading volume. However, due to the specific experimental design they cannot exclude that the results are driven by better-than-average beliefs. Similarly, Michailova and Schmidt (2011) find larger trading volumes in markets consisting

puzzling findings led Barber and Odean (2013) to hypothesize that "this weak link might be partially explained by the current inability to measure miscalibration well." Overall, our results show that there is quite some truth in this hypothesis.

2 Experimental design

2.1 Experimental Procedure

Our experimental design consists of two main parts: The first part involves the IP using ten general knowledge items (see Table A.1 in Appendix B) as usually employed in the literature. To establish comparability of our results, we use the same test as Biais et al. (2005) and Hilton et al. (2011). Participants are instructed to estimate boundaries for these ten general knowledge questions so that their certainty equals the requested level of confidence. We use three treatments with different levels of requested confidence: 30%, 60% and 90%. To support participants' attention to their requested confidence level and their understanding of its implication, we take several precautions. First, we increase the salience of the requested confidence level, expressed in percentage, by clearly highlighting the relevant text from the remaining instructions. Second, we explicitly state how many of the ten true values should on average fall inside of the provided intervals as people may struggle with percentages (Gigerenzer and Hoffrage, 1995). Finally, we explain the intuition of the assignment by a concrete example. Complete instructions of the experiment can be found in Appendix B. After the standard IP procedure, we elicit frequency judgments by requesting participants to estimate how many true items fall inside their intervals.[5] These frequency judgments provide a self-reported proxy for the individual confidence level and allow us to compute an alternative measure of overconfidence that accounts for individual differences in confidence.

In addition, we elicit the perceived degree of difficulty on a five point Likert scale (1=very easy, 5=very difficult) to control for the hard-easy effect (Gigerenzer et al., 1991). Finally, we aim to shed light on the construction of intervals and ask our participants to

of overconfident traders compared to markets with less overconfident traders using an alternative measure of overconfidence based on perceived confidence.

[5]In principle, frequency judgments can be incentivized by means of common incentive compatible mechanisms. Cesarini et al. (2006) compared frequency judgments between two groups of which only one was incentivized and did not find significant differences. In light of this weak evidence of the impact of monetary incentives we refrained from using incentivized frequency judgments.

explain how they generated their interval estimates with a free text question.[6]

In the second part we elicit several behavioral characteristics to analyze whether they can explain individual differences in the IP overconfidence measure. We measure risk- and loss aversion parameters of participants using incentivized binary choice questions similar to Abdellaoui et al. (2008). To test cognitive skills we let participants perform an adapted version of the Cognitive Reflection Test (CRT) of Frederick (2005). Since the IP requires that participants report intervals of their subjective probability distributions for non-random values, we further elicit risk literacy with the Berlin Numeracy Test (BNT) (Cokely et al., 2012). Finally, we assess the degree to which participants avoid the use of wide intervals using a methodology borrowed from Yaniv and Foster (1995). A detailed description of the elicitation methodologies is provided in Appendix B.

2.2 Participants and Incentives

We conducted two online surveys in December 2014 and January 2015 with a total of 300 participants. One survey involves 151 participants on the crowd-sourcing platform Amazon Mechanical Turk (MTurk). The survey instructions were in English using imperial units for the IP questions. MTurk participants received a participation fee of $1.80. To provide additional robustness for our results, we recruited 149 participants from the University of Zurich in a second survey. Invitations for this survey were sent via email to a bachelor level finance class and two master level finance and economics classes. The survey instructions were either in English or German depending on the language of the course that we recruited from. Students were asked to provide IP items in metric units. In each survey, five participants were randomly chosen for a bonus payment to incentivize their risky choices in Part 2 of the survey.[7]

To ensure that our data come from participants who completed all tasks properly and without interruptions, we only include those observations for which the completion time was more than five and less than sixty minutes.[8] These constraints reduce the number of

[6] Additionally, the survey includes two questions for which we do not report the results in this study. We ask participants to estimate the number of hits of the average participant and to provide one-year interval forecasts for the level of either the Dow Jones Industrial Average or the Swiss Market Index depending on the participant sample.

[7] The average payments amounted to 46.8 CHF and $ 22.4 in the Student and MTurk sample, respectively. We deliberately chose larger stakes in the Student sample since a pilot study indicated longer completion times for students. At the time of the study, 46.8 CHF were worth $ 31.43 in the U.S. considering the purchasing power parity level provided by OECD (`http://stats.oecd.org`).

[8] We impose the lower bound time constrained based on experience from a pilot study to exclude MTurk participants that click-through the survey to maximize their hourly wage without carefully reading the

participants to 276 (MTurk sample: 139; Student sample: 137). The fraction of women is 40.3% and 30.7% in MTurk and Student sample, respectively. The two samples resemble typical differences between students and MTurk participants (Goodman et al., 2013): Participants in the MTurk sample have various educational backgrounds and the age distribution is more heterogeneous (between 18 years and 74 years) compared to the younger students (between 18 years and 30 years). Moreover, the median time that students spent on the survey (23:22min) is more than twice as long as in the MTurk sample (9:54min).

Sample	Requested Conf.	Min.	Q1	Median	Mean	Q3	Max.
MTurk	30%	0	20	30	31.3	40	80
MTurk	60%	0	20	30	30.7	40	80
MTurk	90%	0	20	40	40.0	50	100
Biais et al. (2005)	90%	0	20	30	36.0	50	100

Table 1: This table reports summary statistics of the hit rate distributions of all treatments in the MTurk sample. The second column displays the requested confidence level. The columns Q1 and Q3 refer to the first and third quartiles, respectively. All figures are in percentages. This table also reports the hit rate distribution found in Biais et al. (2005).

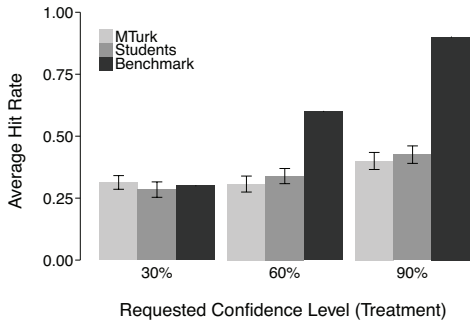
3 Results

The two participant samples yield very coherent results. In this section, we therefore present findings of the MTurk sample and discuss only those results of the student sample that are qualitatively different. A detailed analysis of our findings in the student sample is provided in Appendix A.

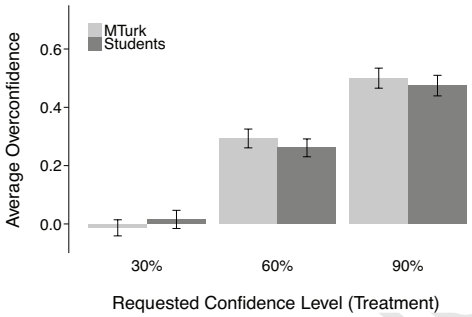
3.1 IP overconfidence measure

Generally, we find that hit rates are very low with a global average of 34.2%. This is in line with previous findings (e.g. Biais et al., 2005; Glaser and Weber, 2007; Hilton et al., 2011). The Cronbach alpha is 0.64 which indicates reasonable internal consistency of the IP. Table 1 shows summary statistics of the hit rate distributions for all three treatments, i.e. requested confidence levels of 30%, 60% and 90%. To rule out that our results are driven by the limited control in online experiments or the potentially low qualification or motivation of MTurk participants, we compare the hit rate distributions of the 90% treatment with

instructions. The upper bound constraint should prevent interruptions. Employing other time constraints (e.g. between 7 and 30 minutes) does not qualitatively change our results. We further exclude two participants who admitted cheating in their full text explanations.



(a) Average hit rates with standard error bars and perfect calibration benchmark grouped by participant samples and treatments.



(b) Average overconfidence based on hit rates with standard error bars grouped by participant samples and treatments.

Figure 1: Hit rates and the IP measure of overconfidence.

the findings of Biais et al. (2005) who solicit master level students to do the same IP with a 90% requested confidence level (and the same IP questions) in a laboratory setting. It turns out that the two distributions are strikingly similar, the MTurk sample even achieves slightly larger hit rates. Hence, our IP results do not seem specific to an online experiment with MTurk participants.

Despite the extreme variation in the requested confidence level from 30% to 90%, average hit rates and hit rate distributions of the three treatments appear almost identical. Figure 1a illustrates the average hit rate of each treatment in comparison to its respective perfect calibration benchmark. In all treatments the hit rates are very similar and seem virtually unaffected by the requested confidence level. This observation is supported in a Kruskal-Wallis rank sum test which does not yield statistically significant differences in hit rates between the three treatments, although just marginally not ($p = 0.10$). Hence, the participants show hardly any reaction to the requested confidence level, which confirms the findings of Teigen and Jørgensen (2005).

Figure 1b illustrates what this result implies for the level of overconfidence in the three treatment groups. Participants in the commonly used 90% and the 60% treatment are overconfident on average, whereas participants in the 30% treatment appear almost perfectly calibrated. Since participants fail to adjust their interval estimates, the IP yields degrees of overconfidence that largely depend on the arbitrary choice of the requested confidence level by the experimenter. As a consequence, it appears difficult to draw conclusions from the IP about the true magnitude of overconfidence at the group level.

We analyze the extent to which behavioral characteristics are able to explain the

Hit rate regressions on behavioral factors

Confidence	CRT	BNT	RA	LA	WIA	Difficulty	Gender	Intercept	Adj. R^2	# obs.
				Panel A: MTurk sample						
0.114	0.131*		−0.033*	−0.007	−0.046	−0.060*	0.001	0.566**	0.101	139
(0.073)	(0.050)		(0.015)	(0.010)	(0.064)	(0.028)	(0.037)	(0.151)		
0.120		0.133*	−0.035*	−0.005	−0.049	−0.071*	0.011	0.651**	0.088	139
(0.074)		(0.061)	(0.015)	(0.010)	(0.065)	(0.028)	(0.037)	(0.149)		
				Panel B: Student sample						
0.220**	0.067		-0.005	0.012	-0.099	0.004	0.029	0.158	0.052	137
(0.079)	(0.068)		(0.024)	(0.013)	(0.071)	(0.029)	(0.047)	(0.141)		
0.205*		0.041	-0.005	0.012	-0.091	0.010	0.041	0.163	0.049	137
(0.079)		(0.056)	(0.024)	(0.013)	(0.070)	(0.028)	(0.044)	(0.142)		

Significance levels: ** : $p < 0.01$; * : $p < 0.05$.

Table 2: This table reports OLS regression coefficients and the respective standard errors in parenthesis of the regression model $HR_i = \alpha + \beta_1 Confidence_i + \beta_2 CA_i + \beta_3 RA_i + \beta_4 LA_i + \beta_5 WIA_i + \beta_6 Difficulty_i + \beta_7 Gender_i + \epsilon_i$ where HR_i is the hit rate, and CA_i is one of the two measures of cognitive abilities: CRT_i or BNT_i. Panel A and B show results of the MTurk and student sample, respectively. The independent variables are as follows: Confidence indicates the requested confidence level; CRT is the ratio of correct answers in the Cognitive Reflection Test; BNT is the ratio of correct answers in the Berlin Numeracy Test; RA is the estimated parameter of risk aversion; LA is the estimated parameter of loss aversion; WIA is ratio of narrow intervals chosen in the wide interval aversion test; Difficulty is the perceived level of difficulty of the question set; Gender is a dummy variable equal to one if the observation stems from a male participant. A detailed explanation of the construction of the independent variables can be found in Appendix B.

between-subject variation in the IP overconfidence measure using OLS regressions with hit rate as dependent variable. To that end, we pool the data of the three treatments and use the requested confidence level as an independent variable. In all regressions we control for gender and the perceived difficulty of the question set. The results are shown in Table 2. We find that participants in the MTurk sample with better cognitive abilities and better numeracy skills achieve larger hit rates and therefore appear significantly less overconfident ($p < 0.05$). This is an interesting finding as it might explain some puzzling results in which experimenters try to link overconfidence to individual economic behavior, e.g. the finding that overconfident investors are less profitable but do not trade more. In line with the results above the requested confidence level does not affect hit rates. Interestingly, we find no gender effects but a significant hard-easy effect indicating a negative relation between hit rates and the perceived difficulty of the question set ($p < 0.05$). Finally, the regression results suggest that a larger degree of risk aversion is associated with lower hit rates ($p < 0.05$).

The regression results in the student sample differ in two main aspects. First, we find that none of the coefficients of the behavioral characteristics is significant. A potential

reason for the insignificant coefficients of the CRT score and the coefficient of risk aversion is the lack of variation in these measures. In the student sample more than 79% (49%) of the students are able to correctly answer at least two (all) of the questions in the CRT compared to 62% (31%) in the MTurk sample. Likewise, the variation in risk aversion is smaller in the student sample ($sd = 0.85$) than in the MTurk sample ($sd = 1.19$). Possibly more sophisticated tests would have led to more pronounced results and cognitive abilities might still be linked with interval widths. The lack of variation cannot explain different results in the BNT score coefficient. But since students take significantly more time to finish the survey one might argue that the ability of the CRT and BNT to discriminate the students by their cognitive abilities is limited. The second difference is that students are slightly more sensitive to the requested confidence level than MTurk participants. The coefficient of the requested confidence level is significant (in one specification at the 1% level). Nevertheless, the size of the effect is in line with the limited sensitivity found in the hit rate results of the student sample (see Appendix A).

Altogether, the regression results are mixed, but suggest that the precision of the IP overconfidence measure might be impaired by additional individual characteristics such as cognitive abilities and risk preferences.

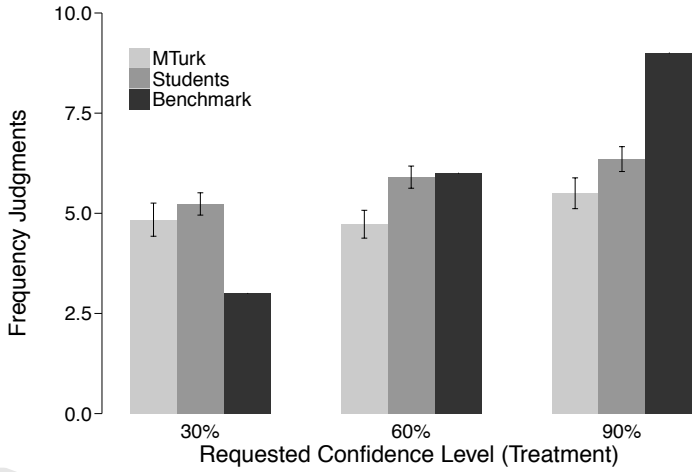


Figure 2: Average frequency judgments with standard error bars grouped by treatment. The dark grey bars indicate the respective benchmark levels of confidence that is assumed by the IP.

3.2 Frequency overconfidence measure

We very clearly instructed participants that they should expect to achieve a hit rate that equals the requested confidence level. Nevertheless, only 12.3% of the participants'

Distribution of frequency judgments

		Bins:	0	1	2	3	4	5	6	7	8	9	10	Total
Panel A: MTurk sample														
Treatments:	30%		3	0	7	6	5	6	4	5	3	2	3	44
	60%		2	1	5	5	5	13	3	4	4	1	1	44
	90%		1	3	4	6	4	8	5	2	8	8	1	50
Panel B: Student sample														
Treatments:	30%		0	0	1	12	4	8	4	10	6	0	0	45
	60%		0	0	3	2	2	4	18	4	7	0	1	41
	90%		2	0	1	2	2	10	6	12	7	8	1	51

Table 3: This table lists for all possible values of frequency judgments the number of participants that chose the respective value.

frequency judgments were exactly in line with the instructions. In all treatments, the average frequency judgments are very close to five and range only from 4.73 in the 60% treatment to 5.50 in the 90% treatment (see Figure 2). Using a Kruskal-Wallis test we cannot reject the null hypothesis that the location parameters of the distributions of frequency judgments in the three treatments are the same. Moreover, we find large heterogeneity in frequency judgments between subjects in all treatments covering the whole range from zero to ten (see Table 3). The standard deviation is between 2.31 and 2.74. In line with this, just 3.6% of the participants mention the requested confidence level in the free text explanations of how they constructed the range estimates. Students appear slightly more sensitive to the instructions than MTurk workers (see Figure 2). However, the main results are qualitively in line with the MTurk sample: we find large variations in frequency judgments in all treatments and only the difference in frequency judgments between the 30% and 90% treatment is statistically significant (see Appendix A.2). The large amount of heterogeneity in frequency judgments indicates that there exist substantial differences in confidence even among participants of the same treatment. Further, the missing treatment effect shows that the requested confidence level hardly affects perceived confidence levels, if at all, even though the relation is explicitly stressed in the instructions. This is in stark contrast to the implicit assumption of the IP.

In the following, we calculate overconfidence based on perceived confidence (frequency overconfidence) as the difference between relative frequency judgments and achieved hit rates for each participant. The resulting distributions of frequency overconfidence are illustrated as boxplots in Figure 3. In line with previous findings (Cesarini et al., 2006; Winman et al., 2004), frequency judgments yield moderate overconfidence. The average level of frequency overconfidence is nearly identical in all treatments at roughly 16%. In contrast to

the IP measure, the differences in overconfidence based on frequency judgments are not statistically significant between the three treatments in a Kruskal-Wallis test. Consequently, frequency overconfidence turns out to be coherent throughout all treatments. Analogous

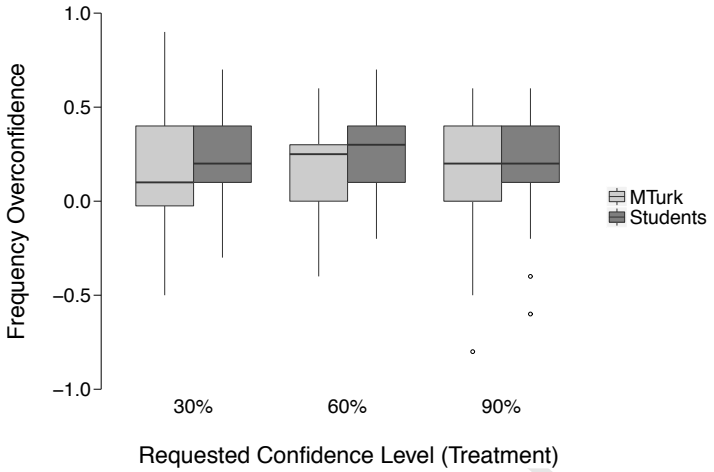


Figure 3: Boxplot of frequency overconfidence grouped by treatments. The edges of the boxes indicate the first and third quartile. The upper and lower whiskers extend from the hinge to the highest and lowest value that is within 1.5 times of the inter quartile range of the hinge, respectively.

to the analysis of the IP measure we run regressions to test if frequency overconfidence is associated with individual characteristics. The regression results (see Table 4) confirm that frequency overconfidence is not affected by requested confidence level. In contrast to the IP measure, frequency overconfidence does not correlate with cognitive abilities. In the MTurk sample larger degrees of frequency overconfidence are positively associated with wide-interval aversion. Moreover, we find some evidence that male participants are more overconfident. Similar to the hit rate regressions we find no significant coefficients in the student sample regressions (see Panel B of Table 4).

It remains the question to what extent the two measures are consistent in their ranking of individual overconfidence. A priori one should expect that the two measures are perfectly consistent since both are designed to measure individual overconfidence. The observed heterogeneity in individual confidence, however, suggests that the IP measure ranks people differently because it is based on a homogeneous confidence level. Additionally, the differences in the regression results indicate limits of the consistency between the two measures. To quantify and analyze the consistency of the two measures we compute the rank correlation coefficient Kendall's τ_b. We do the calculation for each treatment separately due to the treatment effect on the IP overconfidence measure. In the MTurk sample, we find that the correlation ranges between 0.30 (30% and 90% treatment) to 0.42 (60%

Frequency overconfidence regressions on behavioral factors

Confidence	CRT	BNT	RA	LA	WIA	Difficulty	Gender	Intercept	Adj. R^2	# obs.
				Panel A: MTurk sample						
−0.033	0.027		0.017	0.002	0.290**	−0.062	0.095*	0.250	0.109	138
(0.093)	(0.064)		(0.019)	(0.012)	(0.081)	(0.036)	(0.047)	(0.191)		
−0.025		0.113	0.017	0.002	0.286**	−0.068	0.089	0.267	0.123	138
(0.092)		(0.076)	(0.019)	(0.012)	(0.081)	(0.035)	(0.046)	(0.185)		
				Panel B: Student sample						
0.005	0.033		0.044	−0.015	0.093	−0.018	0.058	0.141	0.015	137
(0.082)	(0.070)		(0.025)	(0.013)	(0.073)	(0.030)	(0.048)	(0.146)		
-0.006		0.049	0.046	-0.016	0.094	-0.015	0.059	0.128	0.018	137
(0.082)		(0.058)	(0.025)	(0.013)	(0.072)	(0.029)	(0.045)	(0.146)		

Significance levels: ** : $p < 0.01$: * : $p < 0.05$

Table 4: This table reports OLS regression coefficients and the respective standard errors in parenthesis of the regression model $fOC_i = \alpha + \beta_1 Confidence_i + \beta_2 CA_i + \beta_3 RA_i + \beta_4 LA_i + \beta_5 WIA_i + \beta_6 Difficulty_i + \beta_7 Gender_i + \epsilon_i$ where fOC_i is frequency overconfidence, and CA_i is one of the two measures of cognitive abilities: CRT_i or BNT_i. Panel A and B show results of the MTurk and student sample, respectively. The independent variables are as follows: Confidence indicates the requested confidence level; CRT is the ratio of correct answers in the Cognitive Reflection Test; BNT is the ratio of correct answers in the Berlin Numeracy Test; RA is the estimated parameter of risk aversion; LA is the estimated parameter of loss aversion; WIA is ratio of narrow intervals chosen in the wide interval aversion test; Difficulty is the perceived level of difficulty of the question set; Gender is a dummy variable equal to one if the observation stems from a male participant. A detailed explanation of the construction of the independent variables can be found in Appendix B.

treatment). The correlation coefficients in the student sample are slightly larger and range between 0.41 (90% treatment) and 0.52 (60% treatment). The positive relation of the two measures is significant at the 1% level in all treatments of the student sample and in the 60% treatment of the MTurk sample (in the 30% and 90% treatment of the MTurk sample: $p < 0.05$). The moderate size of the coefficients, however, indicates that the consistency is limited because the two measures disagree in their overconfidence ranking in a considerable number of cases. The level of correlation is at least so low that it will likely effect results on the individual level in studies that analyze the relation between overconfidence and economic behavior. Thus, taking into account the individual confidence level of participants not only reduces aggregate overconfidence but also changes the individual ranking of overconfidence compared to the IP measure. Given the insensitivity of participants to their requested confidence level we conclude that overconfidence based on frequency judgments provides a more accurate measure of individual overconfidence.

4 Conclusion

Judgmental overconfidence gives people excessive trust in their own (biased) judgment. For many areas, such as management, politics, or investment advice, it is extremely important to understand how exactly overconfidence affects the economic behavior of individuals. But to answer this ultimately empirical question, a reliable method of measuring overconfidence is needed. Measuring overconfidence in the field is tricky and has to rely on proxies. Researchers hence "escape" to the laboratory. The most widely used lab test is the Interval Production task (IP). Given the importance of overconfidence for many real-world applications, we critically analyze the appropriateness of the IP to measure judgmental overconfidence. In theory, the test provides an elegant way to measure the extent to which someone overestimates the precision of her own knowledge. It is based on the assumption that people adopt the confidence level that is requested by the experimenter. We find, however, the decision makers' stated intervals and perceived confidence are virtually unaffected by the requested confidence level. Additionally, participants display very different levels of confidence in their interval estimates. By ignoring these individual differences in confidence, the IP measure is likely to yield an imprecise overconfidence ranking. These results are in line with the hypothesis of Barber and Odean (2013) that some puzzling results in behavioral finance may be due to the inability to measure individual overconfidence. We also find some evidence that cognitive abilities are correlated with measured overconfidence which might explain why experimental studies find a link between overconfidence and trading success but not trading volume.

Our results suggest that the widely used classical IP measure is to be treated with caution. Instead of specifying the level of confidence we propose to elicit it. One way of doing so is to use frequency judgments as individual confidence levels. The resulting overconfidence measure yields moderate and coherent overconfidence throughout all treatments and participant samples. We find only limited consistency between the two overconfidence measures. As a consequence, overconfidence based on frequency judgments seems to provide a promising measure to be used in experiments that aim to link overconfidence and economic behavior. Even more so in real world situations – outside the laboratory – people should not be expected to internalize prescribed confidence levels. Our results suggest that a better way of eliciting confidence intervals is to first demand a rough range estimate and subsequently ask for the confidence in that interval.

These are important insights, given the often missing link in experimental studies between measured overconfidence and economic behavior. To draw useful conclusion from

experimental studies on overconfidence for the real world, it will be essential to have robust measurements in place. Our findings show limitations of a currently popular method and provide alternative solutions.

Appendix A Results in the student sample

A.1 IP overconfidence measure

Participants in the student sample achieve fairly low hit rates in all treatments (global average: 35.3%). Overall, the results look very similar to the MTurk sample (see Figure 1a). In fact, we find no significant differences in hit rates between the two samples on the treatment level using Mann-Whitney tests. The Cronbach alpha of 0.67 in the student sample indicates the same level of internal consistency as in the MTurk sample. Within the student sample, we observe very similar hit rate distributions of the three treatments.

In a Kruskal-Wallis test we find significant differences ($p = 0.02$) in the median hit rates which is mainly driven by the difference between the 30% and the 90% treatment (Mann-Whitney: $p < 0.01$). Changing the requested confidence level by 30 percentage points, however, has no significant effect on hit rates.[9]

Thus, we find that in comparison to MTurk workers our students show a slightly larger degree of sensitivity to the requested confidence level. However, we need extreme variations in the requested confidence level to achieve statistical significance despite the relatively large number of participants. By and large, students also fail to sufficiently adjust their intervals. This becomes obvious when we compare the obtained levels of overconfidence between the treatments (see Figure 1b). Almost identical to the MTurk sample, the 30% treatment achieves almost perfect calibration, and overconfidence significantly increases with the level of requested confidence.

A.2 Frequency overconfidence measure

Students seem to adjust their frequency judgments nearly as little as the MTurk participants (see Figure 2). Only about a quarter (27.7%) of the frequency judgments is in line with the requested confidence level. Although this figure is larger than for the MTurk sample, the frequency judgments in the student sample exhibit a similar amount of variation. The standard deviation ranges from 1.77 to 2.22 and frequency judgments cover the whole range of possible answers. Likewise, the full text explanations resemble the same patterns as in the MTurk sample with only 7.3% of the students mentioning their requested confidence level. Within the student sample, we find significant differences in frequency judgments between treatments in a Kruskal-Wallis test ($p = 0.02$). This difference

[9] The non-parametric Mann-Whitney test yields p-values of 0.20 and 0.12 for the comparison of the 30% vs. 60% and 60% vs. 90% treatments, respectively.

is primarily driven by the difference between the 30% and 90% treatment (Mann-Whitney: $p < 0.01$) because a change of the requested confidence level by 30 percentage points has no statistically significant effect.[10] This finding is almost identical to the hit rate results: students display a small amount of sensitivity that becomes visible only for extreme variations of the requested confidence level. As a consequence, the effect size is only marginal and frequency judgments, by and large, appear unaffected by the requested confidence level.

The frequency overconfidence is moderate with an average of 23.2% and slightly larger than in the student sample. The distributions of frequency overconfidence look very similar in all treatments (see Figure 3). Indeed, we cannot reject the null hyptothesis in a Kruskal-Wallis test.

Appendix B Details on the Experimental Design

All surveys were created with the online tool SurveyMonkey^{TM} and participants were provided a link to start the survey either via email or through Amazon Mechanical Turk. Figure A.1 shows a screenshot of the landing page of the survey. In this section, the elicitation methods and tasks used in the survey studies are explained in detail. In addition, instructions from the two surveys conducted with English speaking participants are provided for every task. The order of tasks is identical to the order in the survey.

Socio-demographics: In the beginning of the survey participants are asked to provide socio-demographic information about their gender, age, level of education and average household income.

Interval Production Task (IP): We choose the ten general knowledge items used in Biais et al. (2005) for the IP. To prevent order effects the sequence of the items is randomized for each subject. Figure A.2 shows a screenshot of three of the ten general knowledge items of the IP with a requested confidence level of 60%. A list of all ten items is provided in Table A.1. Every participant reads the following instructions:

On this page you will see 10 text descriptions of numerical general knowledge items, e.g. age at death of Martin Luther King. Please provide a lower bound and upper bound estimate for each value.

[10] A Mann-Whitney test between the 60% and 30% (90%) treatment yields a p-value of 0.13 (0.17).

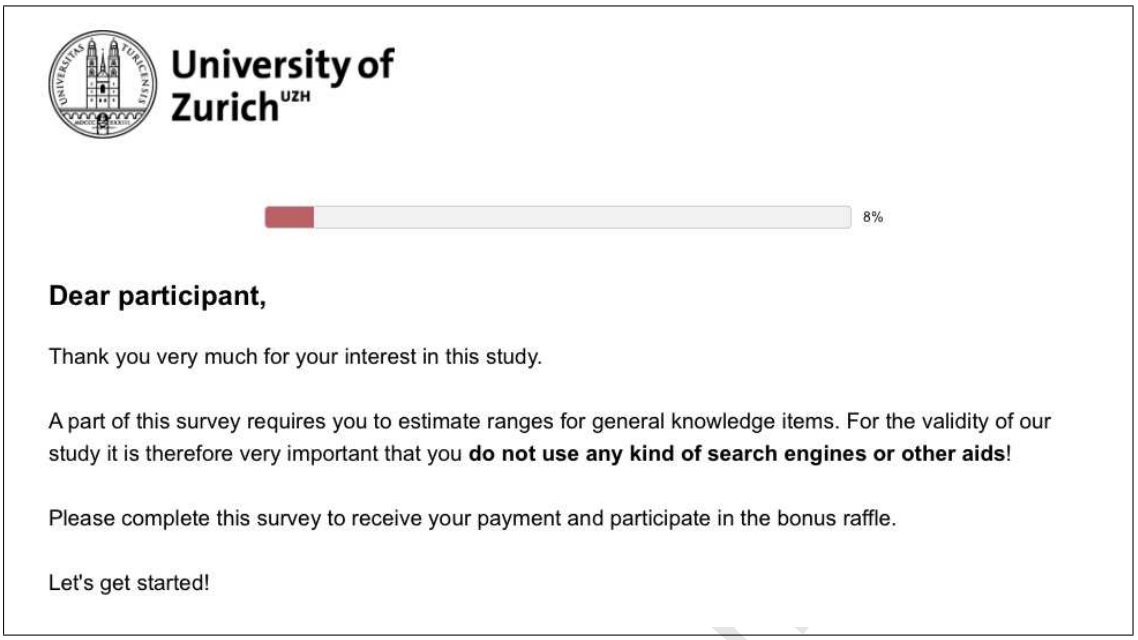


Figure A.1: Screenshot of the landing page of the survey.

You should choose your estimates such that you are 90 % / 60% / 30% certain that the true value lies within your range. This means that on average 9 / 6 / 3 out of 10 true values should fall into your ranges.

Example: Suppose you are 90 % / 60% / 30% certain that the city Berlin has between 2 million and 5 million citizens. Then your answers are: lower bound: 2,000,000; and upper bound 5,000,000.

The 11th item of the IP always involves the stock market prediction task which is optional for all participants. We informed participants about the previous day's closing level of the respective stock index.

Subsequent to the IP, we ask participants the following four questions:

- *How many true values of the first 10 general knowledge items from the previous page do you expect to lie within your provided ranges?*
- *How many true values of the first 10 general knowledge items from the previous page do you expect to lie within the provided ranges of the average Mturker?*

Item
Martin Luther King's age at death
Length of the Nile River
Number of countries that are members of OPEC
Number of books in the Old Testament
Weight of an empty Boeing 747
Year in which the composer Johann Sebastian Bach was born
Average gestation period of an elephant
Diameter of the moon
Air distance from London to Tokyo
Deepest known point in the oceans

Table A.1: General knowledge items of the IP.

- *How difficult did you find the 10 general knowledge items?*
- *(Optional:) How did you come up with your range estimates? (Please write one or two sentences)*

The difficulty question is answered on a 5 Point Likert scale (1 = very difficult; 5 = very easy). The order of answers is counterbalanced such that the scale starts either with very easy or very difficult with probability of 50% each.

Risk- and loss aversion: We elicit parameters for risk- and loss aversion of every participant using choice questions between a sure payoff and a simple binary lottery. The decisions for each domain are displayed on separate a page of the survey. To elicit risk aversion we keep the binary lottery constant for all choices and only increase the sure payoff. We use the number of sure payoff choices as a parameter for risk aversion (RA). The parameter for loss aversion (LA) is obtained analogously. But here, we keep the sure payoff constant at zero and sequentially increase the positive payoff of the risky lottery. Figure A.3 and Figure A.4 show examples of the lotteries for the elicitation of risk- and loss aversion, respectively. The instructions for the risk- and loss aversion elicitation of the survey are as follows:

On the following 3 pages, you have to make ten decisions each between a sure payoff and a coin flip gamble. The coin flip gamble can yield one of two possible payoffs (both with 50% probability). These payoffs can involve gains as well as losses.

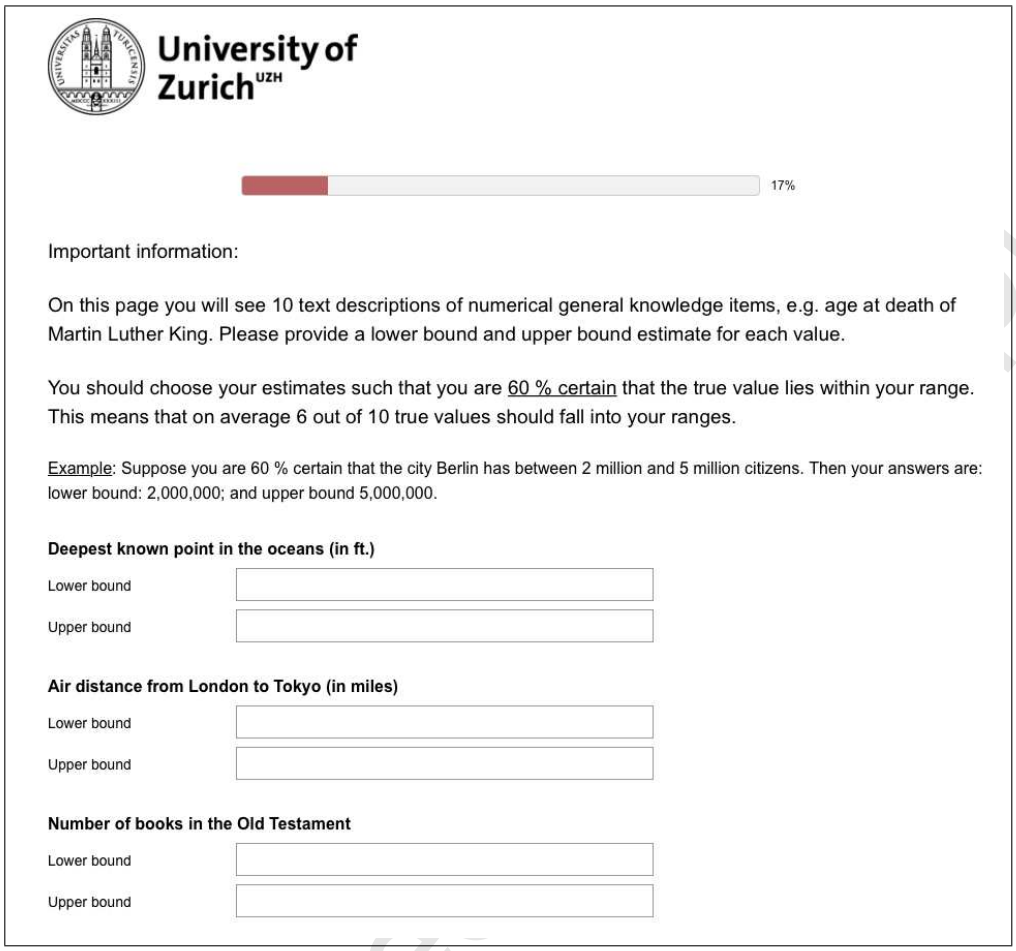


Figure A.2: Screenshot of the IP with a requested confidence level of 60%.

Bonus payment:

You receive an endowment of \$15 (30 CHF). Out of all participants five will be randomly selected to play out one of their ten decisions in real money based on a \$15 (30 CHF) endowment. For these participants one of their ten decisions will be randomly selected and paid out. If you select the sure payoff for the selected decision you will receive this gain plus the \$15 (30 CHF) endowment. If you select the coin-flip gamble for the selected decision one of the possible payoffs will be randomly selected for you with 50% probability each. If it is a gain it will be added to your endowment. If it is a loss it will be subtracted from your endowment.

Please indicate for each choice which of the prospects you would prefer. Bear in

mind that every single decision may be played out in real money!

The last paragraph is shown on the top of each of the three pages.

Cognitive Reflection Test (CRT): We use an amended version of the original cognitive reflection test of Frederick (2005) to prevent people who know the original from using their memorized values.

- *Two iron plates weigh 5.5kg in total. The smaller plate weighs 5kg less than the heavier one. How many kg does the small plate weigh?*
- *3 workers need 3 days to produce 3 guitars. How many days would it take 10 workers for 10 guitars?*
- *On a meadow, there is a patch of flowers. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the meadow, how many days would it take for the patch to cover half of the meadow?*

The test score (CRT) is the ratio of correct answers.

Figure A.3: Risk aversion binary choice.

Berlin Numeracy Test (BNT): We test our participants' skills in probability and statistics using the three question Berlin Numeracy Test of Cokely et al. (2012). The questions are asked in increasing order of difficulty:

1. *Out of 1000 people in a village 500 are a member of the choir. Out of these 500 members 100 are men. Out of the 500 people that don't sing in the choice 300 are men. How big is the probability that a randomly picked man is a member of the choir?*

Figure A.4: Loss aversion binary choice.

2. *In a forest 20% of the mushrooms are red, 50% are brown and 30% are white. A red mushroom is poisonous with probability of 20%. A mushroom that is not red is poisonous with probability of 5%. What is the probability that a poisonous mushroom in the forest is red?*

3. *Imagine you are throwing a biased die (with six sides) 70 times. The probability that the die is twice as high as the probability for each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6?*

The test score (BNT) is the ratio of correct answers.

Wide-interval aversion (WIA): The last part of the survey is designed to obtain a proxy of the aversion towards wide intervals. The test is borrowed from Yaniv and Foster (1995) who argue that people are facing a trade-off between informativeness and accuracy when answering the IP: The wider the interval estimate the more likely it is that the true value falls inside of it (high accuracy), but at the same time the information content of the statement is diluted (low informativeness) and vice versa. They find evidence that people tend to provide narrow intervals due to a preference for high informativeness (Yaniv and Foster, 1997). The elicitation task involves choosing a preferred interval description for a certain value. Each choice is made between two intervals of which one is always wider and includes the true value and the other is narrower but does not include the true value. For all three choices the wide interval is approximately centered around the true value. The choices only differ in the distance of the narrow interval from the true value. Below you find the three choices in increasing order of distance. In the experiment the order of items was randomized for all subjects. The measure of aversion towards wide intervals (WIA) is

the ratio of narrow intervals chosen. The complete task is shown below:

A couple of years ago a researcher had to prepare a presentation. Since he was missing a couple of precise values he asked two colleagues for an estimation of the true values. Now - thanks to the internet and search engines - the researcher is able to look up the precise true values and is able to evaluate the estimates of his colleagues.

Imagine you are this researcher. Please indicate for each value which estimate you think is better.

- *Amount of money spent on education by the US federal government in 1987. The true value is: $22.5 billions. Which estimate do you think is better?* (A:$20 billions to $40 billions; B:$18 billions to $20 billions)
- *Start of the Sino (Chinese) - Japanese war. The true value is: 1894. Which estimate do you think is better?* (A: 1870 to 1890; B: 1875 to 1925)
- *Air distance between Chicago and New York. The true value is: 717 miles. Which estimate do you think is better?* (A: 800 miles to 850 miles; B: 600 miles to 800 miles)

References

Abdellaoui, M., Bleichrodt, H., L'Haridon, O., 2008. A tractable method to measure utility and loss aversion under prospect theory. Journal of Risk and Uncertainty 36, 245–266. doi:10.1007/s11166-008-9039-8.

Ackert, L.F., Church, B.K., Qi, L., 2015. An Experimental Examination of Portfolio Choice. Review of Finance doi:10.1093/rof/rfv036.

Alpert, M., Raiffa, P.A., 1982. A progress report on the training of probability assessors, in: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment under uncertainty: Heuristics and biases. Cambridge Univ. Press., Cambridge, England, pp. 294–305.

Barber, B.M., Odean, T., 2000. Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors. Journal of Finance LV, 773–806. doi:10.1111/0022-1082.00226.

Barber, B.M., Odean, T., 2001. Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment. The Quarterly Journal of Economics 116, 261–292. doi:10.1016/S0169-2070(03)00031-1.

Barber, B.M., Odean, T., 2013. The Behavior of Individual Investors, in: Constantinides, G.M., Harris, M., Stulz, R.M. (Eds.), Handbook of the Economics of Finance. Elsevier. volume 2, Part B. chapter 22, pp. 1533–1570. doi:10.1016/B978-0-44-459406-8.00022-6.

Bazerman, M.H., Moore, D.A., 2013. Judgment in Managerial Decision Making. 8th ed., Wiley, New York, USA.

Ben-David, I., Graham, J.R., Harvey, C.R., 2013. Managerial Miscalibration. Quarterly Journal of Economics 128, 1547–1584. doi:10.1093/qje/qjt023.

Biais, B., Hilton, D., Mazurier, K., Pouget, S., 2005. Judgemental Overconfidence, Self-Monitoring, and Trading Performance in an Experimental Financial Market. The Review of Economic Studies 72, 287–312. doi:10.1111/j.1467-937X.2005.00333.x.

Budescu, D.V., Du, N., 2007. Coherence and Consistency of Investors' Probability Judgments. Management Science 53, 1731–1744. doi:10.1287/mnsc.1070.0727.

Camerer, C., Lovallo, D., 1999. Overconfidence and Excess Entry : An Experimental Approach. American Economic Review 89, 306–318. doi:10.1257/aer.89.1.306.

Cesarini, D., Sandewall, Ö., Johannesson, M., 2006. Confidence interval estimation tasks and the economics of overconfidence. Journal of Economic Behavior & Organization 61, 453–470. doi:10.1016/j.jebo.2004.10.010.

Cokely, E.T., Galesic, M., Schulz, E., Ghazal, S., Garcia-Retamero, R., 2012. Measuring Risk Literacy: The Berlin Numeracy Test. Judgment and Decision Making 7, 25–47.

Deaves, R., Luders, E., Luo, G.Y., 2008. An Experimental Test of the Impact of Overconfidence and Gender on Trading Activity. Review of Finance 13, 555–575. doi:10.1093/rof/rfn023.

Fellner-Röhling, G., Krügel, S., 2014. Judgmental overconfidence and trading activity. Journal of Economic Behavior & Organization 107, 827–842. doi:10.1016/j.jebo.2014.04.016.

Frederick, S., 2005. Cognitive Reflection and Decision Making. Journal of Economic Perspectives 19, 25–42. doi:10.1257/089533005775196732.

Gigerenzer, G., Hoffrage, U., 1995. How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review 102, 684–704. doi:10.1037/0033-295X.102.4.684.

Gigerenzer, G., Hoffrage, U., Kleinbölting, H., 1991. Probabilistic mental models: a Brunswikian theory of confidence. Psychological Review 98, 506–28. doi:10.1037//0033-295X.98.4.506.

Glaser, M., Weber, M., 2007. Overconfidence and trading volume. The Geneva Risk and Insurance Review 32, 1–36. doi:10.1007/s10713-007-0003-3.

Goetzmann, W.N., Kumar, A., 2008. Equity Portfolio Diversification. Review of Finance 12, 433–463. doi:10.1093/rof/rfn005.

Goodman, J.K., Cryder, C.E., Cheema, A., 2013. Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. Journal of Behavioral Decision Making 26, 213–224. doi:10.1002/bdm.1753.

Herz, H., Schunk, D., Zehnder, C., 2014. How do judgmental overconfidence and overoptimism shape innovative activity? Games and Economic Behavior 83, 1–23. doi:10.1016/j.geb.2013.11.001.

Hilton, D., Régner, I., Cabantous, L., Charalambides, L., Vautier, S., 2011. Do Positive Illusions Predict Overconfidence in Judgment? A Test Using Interval Production and Probability Evaluation Measures of Miscalibration. Journal of Behavioral Decision Making 24, 117–139. doi:10.1002/bdm.678.

Johnson, D.D., Tierney, D., 2011. The Rubicon Theory of War: How the Path to Conflict Reaches the Point of No Return. International Security 36, 7–40. doi:10.1162/ISEC_a_00043.

Juslin, P., 1993. An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. European Journal of Cognitive Psychology 5, 55–71. doi:10.1080/09541449308406514.

Juslin, P., 1994. The Overconfidence Phenomenon as a Consequence of Informal Experimenter-Guided Selection of Almanac Items. Organizational Behavior and Human Decision Processes 57, 226–246. doi:10.1006/obhd.1994.1013.

Kirchler, E., 2002. Simultaneous Over- and Underconfidence : Evidence from Experimental Asset Markets. The Journal of Risk and Uncertainty 25, 65–85. doi:10.1023/A:1016319430881.

Klayman, J., Soll, J.B., Gonza, C., 1999. Overconfidence : It Depends on How, What, and Whom You Ask. Organizational Behavior and Human Decision Processes 79, 216–247. doi:10.1006/obhd.1999.2847.

Koriat, A., Lichtenstein, S., Fischhoff, B., 1980. Reasons for Confidence. Journal of Experimental Psychology: Learning, Memory, and Cognition 6, 107–118. doi:10.1037/0278-7393.6.2.107.

Langer, E.J., 1975. The Illusion of Control. Journal of Personality and Social Psychology 32, 311–328. doi:10.1037/0022-3514.32.2.311.

Lichtenstein, S., Fischhoff, B., Phillips, L.D., 1982. Calibration of Probabilities: The State of the Art to 1980, in: Kahneman, D., Slovic, P., Tversky, A. (Eds.), Judgment under uncertainty: Heuristics and biases.

Malmendier, U., Tate, G., 2005. CEO overconfidence and corporate investment. Journal of Finance 60, 2661–2700. doi:10.1111/j.1540-6261.2005.00813.x.

Menkhoff, L., Schmidt, U., Brozynski, T., 2006. The impact of experience on risk taking, overconfidence, and herding of fund managers: Complementary survey evidence. European Economic Review 50, 1753–1766. doi:10.1016/j.euroecorev.2005.08.001.

Michailova, J., Schmidt, U., 2011. Overconfidence and Bubbles in Experimental Asset Markets. Working Paper .

Moore, D.A., Healy, P.J., 2008. The trouble with overconfidence. Psychological Review 115, 502–517. doi:10.1037/0033-295X.115.2.502.

Moore, D.A., Tenney, E.R., Haran, U., 2014. Overprecision in judgment, in: Wu, G., Keren, G. (Eds.), Handbook of Judgment and Decision Making. Wiley, New York, USA, pp. 1–50. doi:10.1002/9781118468333.ch6.

Odean, T., 1998. Volume, Volatility, Price, and Profit When All Traders Are Above Average. Journal of Finance LIII, 1887–1934. doi:10.1111/0022-1082.00078.

Ren, Y., Croson, R., 2013. Overconfidence in Newsvendor Orders: An Experimental Study. Management Science 59, 2502–2517. doi:doi:10.1287/mnsc.2013.1715.

Russo, J.E., Schoemaker, P.J., 1992. Managing Overconfidence. Sloan Management Review 33, 7–17.

Soll, J.B., 1996. Determinants of Overconfidence and Miscalibration: The Roles of Random Error and Ecological Structure. Organizational Behavior and Human Decision Processes 65, 117–137. doi:10.1006/obhd.1996.0011.

Soll, J.B., Klayman, J., 2004. Overconfidence in Interval Estimates. Journal of Experimental Psychology: Learning, Memory, and Cognition 30, 299–314. doi:10.1037/0278-7393.30.2.299.

Statman, M., Thorley, S., Vorkink, K., 2006. Investor Overconfidence and Trading Volume. The Review of Financial Studies 19, 1531–1565. doi:10.1093/rfs/hhj032.

Svenson, O., 1981. Are we all less risky and more skillful than our fellow drivers? Acta Psychologica 47, 143–148. doi:10.1016/0001-6918(81)90005-6.

Teigen, K.H., Jørgensen, M., 2005. When 90% confidence intervals are 50% certain: on the credibility of credible intervals. Applied Cognitive Psychology 19, 455–475. doi:10.1002/acp.1085.

Winman, A., Hansson, P., Juslin, P., 2004. Subjective Probability Intervals: How to Reduce Overconfidence by Interval Evaluation. Journal of Experimental Psychology: Learning, Memory, and Cognition 30, 1167–1175. doi:10.1037/0278-7393.30.6.1167.

Yaniv, I., Foster, D.P., 1995. Graininess of Judgment Under Uncertainty : An Accuracy-Informativeness Trade-Off. Journal of Experimental Psychology: General 124, 424–432. doi:10.1037/0096-3445.124.4.424.

Yaniv, I., Foster, D.P., 1997. Precision and Accuracy of Judgmental Estimation. Journal of Behavioral Decision Making 10, 21–32. doi:10.1002/(SICI)1099-0771(199703)10:1<21::AID-BDM243>3.0.CO;2-G.

Yaniv, I., Schul, Y., 1997. Elimination and Inclusion Procedures in Judgment. Journal of Behavioral Decision Making 10, 211–220. doi:10.1002/(SICI)1099-0771(199709)10:3<211::AID-BDM250>3.0.CO;2-J.

Manuscript: Do We Measure Overconfidence? A Closer Look at the Interval Production Task

Highlights:

- People's perceived confidence is unaffected by the requested confidence level
- The IP overconfidence measure neglects heterogeneity in perceived confidence
- The consistency of the IP and frequency overconfidence measures is limited
- The two overconfidence measures are related to different characteristics
- Our findings might explain the missing link of many experimental studies